REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	New Reprint	-

4. TITLE AND SUBTITLE

A scalable silicon photonic chip-scale optical switch for high performance computing systems

5a. CONTRACT NUMBER
W911NF-13-1-0090

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Runxiang Yu, Stanley Cheung, Yuliang Li, Katsunari Okamoto, Roberto Proietti, Yawei Yin, S. J. B. Yoo

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

University of California - Davis
1850 Research Park Drive
Suite 300
Davis, CA 95618 -6153

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
63311-CS-ACI.2

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

This paper discusses the architecture and provides performance studies of a silicon photonic chip-scale optical switch for scalable interconnect network in high performance computing systems. The proposed switch exploits optical wavelength parallelism and wavelength routing characteristics of an Arrayed Waveguide Grating Router (AWGR) to allow contention resolution in the wavelength domain. Simulation results from a cycle-accurate network simulator indicate that, even with only two transmitter/receiver pairs per node, the switch exhibits lower end-to-end latency and higher throughput at high (>90%) input loads compared with electronic switches. On the

15. SUBJECT TERMS

Optical interconnects; Integrated optoelectronic circuits; Switching; Coupled resonators; Integrated optics devices.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			S. J. Ben Yoo
UU	UU	UU	UU		19b. TELEPHONE NUMBER 530-752-7063

Report Title

A scalable silicon photonic chip-scale optical switch for high performance computing systems

ABSTRACT

This paper discusses the architecture and provides performance studies of a silicon photonic chip-scale optical switch for scalable interconnect network in high performance computing systems. The proposed switch exploits optical wavelength parallelism and wavelength routing characteristics of an Arrayed Waveguide Grating Router (AWGR) to allow contention resolution in the wavelength domain. Simulation results from a cycle-accurate network simulator indicate that, even with only two transmitter/receiver pairs per node, the switch exhibits lower end-to-end latency and higher throughput at high (>90%) input loads compared with electronic switches. On the device integration level, we propose to integrate all the components (ring modulators, photodetectors and AWGR) on a CMOS-compatible silicon photonic platform to ensure a compact, energy efficient and cost-effective device. We successfully demonstrate proof-of-concept routing functions on an 8×8 prototype fabricated using foundry services provided by OpSIS-IME.

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 63311.2-CS-ACI
A scalable silicon photonic chip-scale optical sw...

Block 13: Supplementary Note

Approved for public release; distribution is unlimited.

A scalable silicon photonic chip-scale optical switch for high performance computing systems

Runxiang Yu,[1] Stanley Cheung,[1] Yuliang Li,[1] Katsunari Okamoto,[2] Roberto Proietti,[1] Yawei Yin,[1] and S. J. B. Yoo[1,*]

[1]*Department of Electrical Engineering, University of California, Davis, One shields avenue, Davis, CA, 95616, USA*
[2] *AiDi Corporation, 2-2-4 Takezono, Tsukuba, Ibaraki, 305-0032 Japan*
[*]*sbyoo@ucdavis.edu*

Abstract: This paper discusses the architecture and provides performance studies of a silicon photonic chip-scale optical switch for scalable interconnect network in high performance computing systems. The proposed switch exploits optical wavelength parallelism and wavelength routing characteristics of an Arrayed Waveguide Grating Router (AWGR) to allow contention resolution in the wavelength domain. Simulation results from a cycle-accurate network simulator indicate that, even with only two transmitter/receiver pairs per node, the switch exhibits lower end-to-end latency and higher throughput at high (>90%) input loads compared with electronic switches. On the device integration level, we propose to integrate all the components (ring modulators, photodetectors and AWGR) on a CMOS-compatible silicon photonic platform to ensure a compact, energy efficient and cost-effective device. We successfully demonstrate proof-of-concept routing functions on an 8 × 8 prototype fabricated using foundry services provided by OpSIS-IME.

OCIS codes: (200.4650) Optical interconnects; (250.3140) Integrated optoelectronic circuits; (250.6715) Switching; (230.4555) Coupled resonators; (230.3120) Integrated optics devices.

References and links

1. R. Luijten, W. E. Denzel, R. R. Grzybowski, and R. Hemenway, "Optical interconnection networks: The OSMOSIS project," in Lasers and Electro-Optics Society, 2004. LEOS 2004. The 17th Annual Meeting of the IEEE. 2004.
2. M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in ACM SIGCOMM Computer Communication Review. 2008. ACM.
3. A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "VL2: a scalable and flexible data center network," in ACM SIGCOMM Computer Communication Review. 2009. ACM.
4. B. Jalali and S. Fathpour, "Silicon photonics," J. Lightwave Technol. **24**(12), 4600–4615 (2006).
5. S. T. S. Cheung, B. Guan, S. S. Djordjevic, K. Okamoto, and S. J. B. Yoo, "Low-loss and high contrast silicon-on-insulator (SOI) arrayed waveguide grating," in Lasers and Electro-Optics (CLEO), 2012 Conference on. 2012.
6. P. Cheben, J. H. Schmid, A. Delâge, A. Densmore, S. Janz, B. Lamontagne, J. Lapointe, E. Post, P. Waldron, and D. X. Xu, "A high-resolution silicon-on-insulator arrayed waveguide grating microspectrometer with sub-micrometer aperture waveguides," Opt. Express **15**(5), 2299–2306 (2007).
7. P. Dong, S. Liao, D. Feng, H. Liang, D. Zheng, R. Shafiiha, C.-C. Kung, W. Qian, G. Li, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, "Low Vpp, ultralow-energy, compact, high-speed silicon electro-optic modulator," Opt. Express **17**(25), 22484–22490 (2009).
8. D. Ahn, C. Y. Hong, J. Liu, W. Giziewicz, M. Beals, L. C. Kimerling, J. Michel, J. Chen, and F. X. Kärtner, "High performance, waveguide integrated Ge photodetectors," Opt. Express **15**(7), 3916–3921 (2007).
9. H. Park, A. W. Fang, O. Cohen, R. Jones, M. J. Paniccia, and J. E. Bowers, "A Hybrid AlGaInAs-Silicon Evanescent Amplifier," IEEE Photon. Technol. Lett. **19**(4), 230–232 (2007).
10. A. W. Fang, H. Park, O. Cohen, R. Jones, M. J. Paniccia, and J. E. Bowers, "Electrically pumped hybrid AlGaInAs-silicon evanescent laser," Opt. Express **14**(20), 9203–9210 (2006).
11. X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: A scalable optical switch for datacenters," in Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems. 2010. ACM.

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

12. K. Xi, Y.-H. Kao, M. Yang, and H. Chao, "Petabit optical switch for data center networks," Polytechnic Institute of New York University, New York, Tech. Rep.(2010).
13. J. Gripp, J. Simsarian, J. LeGrange, P. Bernasconi, and D. Neilson, "Photonic terabit routers: the IRIS project," in Optical Fiber Communication Conference. 2010. Optical Society of America.
14. H. Yang and S. J. B. Yoo, "Combined input and output all-optical variable buffered switch architecture for future optical routers," IEEE Photon. Technol. Lett. **17**(6), 1292–1294 (2005).
15. OpSIS, Available from: http://opsisfoundry.org/.
16. K. Okamoto, T. Hasegawa, O. Ishida, A. Himeno, and Y. Ohmori, "32× 32 arrayed-waveguide grating multiplexer with uniform loss and cyclic frequency characteristics," Electron. Lett. **33**(22), 1865–1866 (1997).
17. R. Proietti, Y. Yawei, Y. Runxiang, C. Nitta, V. Akella, and S. J. B. Yoo, "An All-Optical Token Technique Enabling a Fully-Distributed Control Plane in AWGR-Based Optical Interconnects," J. Lightwave Technol. **31**(3), 414–422 (2013).
18. Y. Yin, R. Proietti, C. J. Nitta, V. Akella, C. Mineo, and S. J. B. Yoo, "AWGR-based all-to-all optical interconnects using limited number of wavelengths," in Optical Interconnects Conference, 2013 IEEE. 2013.
19. Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson, "12.5 Gbit/s carrier-injection-based silicon micro-ring silicon modulators," Opt. Express **15**(2), 430–436 (2007).
20. H. L. R. Lira, S. Manipatruni, and M. Lipson, "Broadband hitless silicon electro-optic switch for on-chip optical networks," Opt. Express **17**(25), 22271–22280 (2009).
21. A. Biberman, "Silicon Photonics for High-Performance Interconnection Networks," 2011, PhD dissertation, Columbia University.
22. W. Bogaerts, P. Dumon, D. V. Thourhout, D. Taillaert, P. Jaenen, J. Wouters, S. Beckx, V. Wiaux, and R. G. Baets, "Compact Wavelength-Selective Functions in Silicon-on-Insulator Photonic Wires," IEEE J. Sel. Top. Quantum Electron. **12**(6), 1394–1401 (2006).
23. K. Duk-Jun, L. Jong-Moo, S. Jung-Ho, P. Junghyung, and K. Gyungock, "Crosstalk reduction of silicon nanowire AWG with shallow-etched grating arms," in Group IV Photonics, 2008 5th IEEE International Conference on. 2008.
24. H. Yamada, K. Takada, Y. Inoue, K. Okamoto, and S. Mitachi, "Low-crosstalk arrayed-waveguide grating multi/demultiplexer with phase compensating plate," Electron. Lett. **33**(20), 1698–1699 (1997).
25. F. M. Soares, J. H. Baek, N. K. Fontaine, X. Zhou, Y. Wang, R. P. Scott, J. P. Heritage, C. Junesand, S. Lourdudoss, K. Y. Liou, R. A. Hamm, W. Wang, B. Patel, S. Vatanapradit, L. A. Gruezke, W. T. Tsang, and S. J. B. Yoo, "Monolithically integrated InP wafer-scale 100-channel 10-GHz AWG and Michelson interferometers for 1-THz-bandwidth optical arbitrary waveform generation," in Optical Fiber Communication (OFC), collocated National Fiber Optic Engineers Conference, 2010 Conference on (OFC/NFOEC). 2010.
26. B. G. Lee, A. Biberman, D. Po, M. Lipson, and K. Bergman, "All-Optical Comb Switch for Multiwavelength Message Routing in Silicon Photonic Networks," IEEE Photon. Technol. Lett. **20**(10), 767–769 (2008).

1. Introduction

Scalable, low latency, and high-throughput interconnection is essential for future high performance computing (HPC) applications [1]. Interconnect networks based on electronic multistage topologies (e.g. Fat-Tree, CLOS, Torus, Flattened Butterfly [2, 3]) result in large latencies, due to the multi-hop nature of these networks and high power consumption in the buffers and the switch fabric. It is increasingly difficult to meet high bandwidth and low latency communications using conventional electrical switches. On the other hand, integrated optics may enable the continued scaling of capacity required by future HPC systems. Silicon photonics is now the most active discipline within the field of integrated optics due to its compatibility with the mature silicon IC manufacturing. Other motivations include the availability of high quality high index contrast silicon-on-insulator (SOI) wafer to enable the scaling of photonic devices to the hundreds of nanometer level and excellent material properties such as high thermal conductivity, high optical damage threshold and high optical nonlinearities [4]. Recent advances in key components, such as high-port-count low-loss silicon AWG [5] and AWGR [6], Si ring modulators [7], high-responsivity epitaxial Germanium (Ge) photodetectors (PD) [8], hybrid semiconductor optical amplifiers (SOA) [9] and laser sources [10], are paving the way for a disruptive step in device integration for large chip-scale optical switch systems.

Among all the proposed and existing optical interconnect architectures for HPC and datacenters, AWGR based solutions have drawn strong attention due to its dense interconnectivity and unique wavelength routing capability. For example LIONS, (previously named as DOS) [11], Petabit [12], and IRIS [13] are all based on AWGR and Tunable Wavelength Converters (TWC). They benefit from the high capacity offered by Wavelength

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

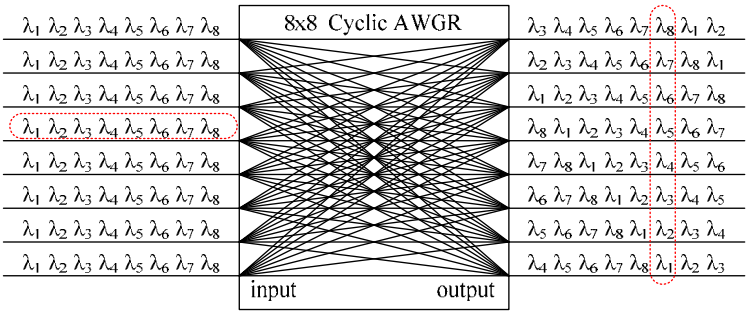
Division Multiplexing (WDM). Furthermore, multiple WDM channels on one output can be used as multiple concurrent channels to avoid head-of-line blocking [14], which results in lower latency and higher throughput. In particular, LIONS uses single fixed wavelength transmitter per node with SOA-MZI based tunable wavelength converters placed at AWGR inputs to route the traffic to the desired AWGR output ports. The 1-by-k DEMUX and k parallel receivers at each output node accommodates up to k concurrent packets using k different wavelengths, which greatly reduces the contention probability and the average end-to-end latency. The contented packets enter an electrical shared loopback buffer and re-enter the AWGR through a dedicated AWGR input/output port pair. A centralized electrical control plane handles all the contention and packet retransmission [11]. Note that, the above LION switch architecture was designed for rack-to-rack or cluster to cluster application, while this paper discusses new AWGR-based switch architecture for on-chip communication in HPC systems. In this case, thanks to the very short distance between nodes and switch, the nodes (processors) communicate with the centralized controller directly in the electrical domain, and the packets, stored in the input queues, are transmitted only upon the node requests are acknowledged and the grants are received. So, this on-chip architecture does not require any electrical loopback buffer and wavelength converters at the AWGR inputs. The tunable lasers can be used directly at the node TXs. In particular, multiple TXs per node can be used to form multiple transmitter/receiver pairs on each connecting nodes. Simulation results show that, even with only two transmitter/receiver pairs, end-to-end latency and throughput is significantly improved compared to its electrical counterpart at high (>90%) input load. In addition, we observe zero packet loss even at 100% input load under the simulated scenarios. In terms of photonic device implementation, we propose to use silicon ring modulators with a broadcasted optical comb source to replace the SOA-MZI TWCs on the transmitter side, and use ring resonators as DEMUXs on the receiver side. The main building blocks (AWGRs, ring resonators and Ge PDs) are all available on the Silicon-On-Insulator (SOI) platform, which results in a compact and cost-effective device. Finally, we present a prototype based on 8 × 8 200-GHz spaced AWGR with four transmitter/receiver pairs on each node. The footprint of the fabricated device is 1.2 mm by 2.4 mm using standard microelectronic foundry service offered by OpSIS-IME [15].

We organize the remainder of the paper as follows: Section 2 describes the proposed scalable optical interconnect architecture based on AWGR with ring resonators and the Ge photodetectors on a SOI platform. Section 3 presents the performance study of the proposed switch using a clock-cycle-accurate architecture-level simulator. Section 4 describes the design of the 8 × 8 switch prototype and presents experimental demonstrations of successful routing functions on the fabricated device by OpSIS-IME. Section 5 concludes the paper.

2. AWGR based optical interconnects with multiple transmitters/receivers at each node

Wavelength division multiplexing (WDM) technology allows for the frequency domain parallelism. Meanwhile, AWGR allows for the multiplexed wavelengths in the waveguides to be separated and cross-connected. As shown in Fig. 1, N nodes (N = 8 in this example) respectively connected to the N input ports of an AWGR can use N wavelengths to reach different output ports simultaneously without interfering with each other. The cyclic frequency feature [16] guarantees the same set of wavelengths can be used at each node. In principle, the single passive AWGR-based all-to-all interconnections of N nodes in a star topology provides the densest communication pattern that can be implemented in a computing network provided that an $N \times 1$ optical multiplexer ($1 \times N$ de-multiplexer) and N transmitters (receivers) are available for each AWGR input (output) ports. This configuration requires N^2 transmitter/receiver pairs in total, which does not scale. Alternatively, AWGR with fixed k_t (k_t<N) transmitters and k_r (k_r<N) receivers at each interconnecting node can potentially provide a scalable and efficient solution [11].

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013



(a)

	Output 1	Output 2	Output 3	Output 4	Output 5	Output 6	Output 7	Output 8
Input 1	λ_3	λ_2	λ_1	λ_8	λ_7	λ_6	λ_5	λ_4
Input 2	λ_2	λ_1	λ_8	λ_7	λ_6	λ_5	λ_4	λ_3
Input 3	λ_1	λ_8	λ_7	λ_6	λ_5	λ_4	λ_3	λ_2
Input 4	λ_8	λ_7	λ_6	λ_5	λ_4	λ_3	λ_2	λ_1
Input 5	λ_7	λ_6	λ_5	λ_4	λ_3	λ_2	λ_1	λ_8
Input 6	λ_6	λ_5	λ_4	λ_3	λ_2	λ_1	λ_8	λ_7
Input 7	λ_5	λ_4	λ_3	λ_2	λ_1	λ_8	λ_7	λ_6
Input 8	λ_4	λ_3	λ_2	λ_1	λ_8	λ_7	λ_6	λ_5

(b)

Fig. 1. (a) The routing property and (b) the routing table of an 8 × 8 cyclic-frequency AWGR.

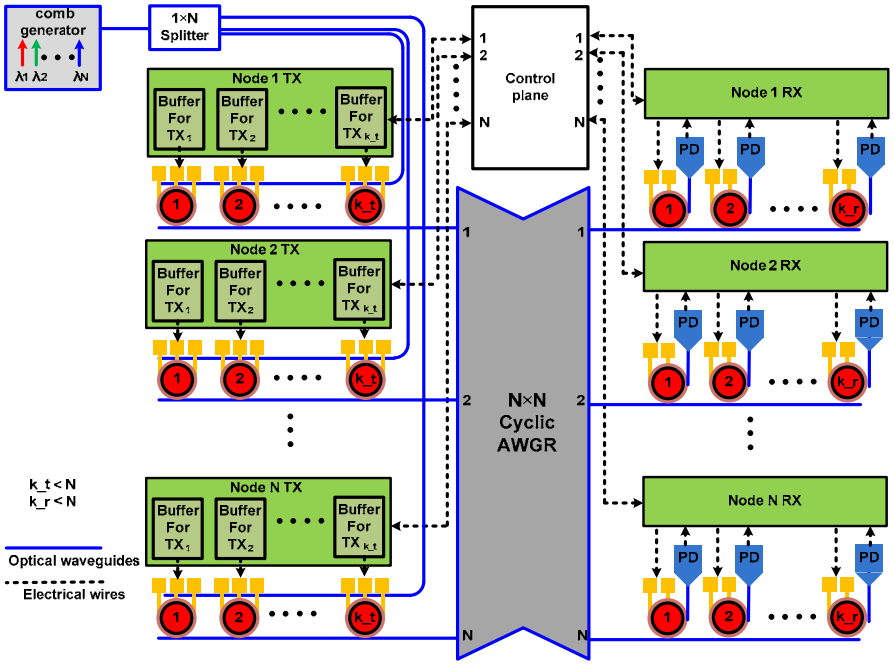


Fig. 2. The proposed interconnect architecture for chip-scale high performance computing.

Figure 2 shows the proposed chip-scale interconnect architecture. We assume a centralized control plane here for simplicity, which can be potentially replaced by a distributed one using the all-optical TOKEN technique [17]. Each node has a transmitter array that uses k_t ring modulators to generate the data packets. An off-chip comb generator provides the N wavelengths required by the cyclic frequency AWGR for wavelength routing. We route the waveguides in such a way that only selected comb lines enter the AWGR to avoid crosstalk from the unused ones. Another solution is to assign a fast tunable laser for each ring

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

modulator. We choose silicon ring modulator due to its compactness, wavelength selectivity and energy efficiency [7]. The rings at the input side serve as both optical modulator and wavelength MUX. As a result, they have two sets of the control electrodes. The low speed pads are for aligning the ring resonances with the AWGR passbands while we apply high-speed RF signals on the Ground-Signal or Ground-Signal-Ground pads for data modulation. Only low speed pads are required for the DEMUX rings on the output bus waveguides. The receiver reads the information on the de-multiplexed wavelength after Optical-to-Electrical (O/E) conversion. Since we have multiple rings on both input and output bus waveguides, one node can communicate with multiple other nodes simultaneously. We restrict the number of rings on each bus waveguide to a fix number k_t and k_r. Other than the drivers and buffers for the modulators and the detectors, all the electronic components can work at a much lower speed than the line rate, so the proposed switch can be potentially more power efficient than the conventional electronic switches.

The scalability of the proposed switch depends on the scalability of AWGR. In theory, fabrication of large-port-count AWGRs is possible, but limiting factors, such as difficulties in accurate wavelength registration and high crosstalk due to dense channel spacing and large number of channel, prevent such system from being deployed in a large scale. However, it is feasible to use small AWGRs with a fewer number of wavelengths while supporting the same connectivity as large-port-count AWGR [18]. Interconnecting N nodes is possible by using W ($W < N$) wavelengths and $W \times W$ AWGRs. Meanwhile, ring FSR should be larger than $N \times$ AWGR channel spacing in a single AWGR configuration, so ring resonance only aligns with one of the AWGR passbands. Assume a ring resonator with a 5-um radius [19] (approx 2.4-THz FSR), the switch can easily accommodate 32 (48 maximum) wavelegnth channels with a 50-GHz channel spacing. Based on the above analysis, larger scale switch can be built from 32×32 AWGRs [6] and ring resonators with 5-um radii.

3. Performance study of the proposed interconnect

We develop a clock-cycle-accurate architecture-level simulator and simulate the proposed interconnect switch with 8 nodes and 64 nodes. For simplicity, we only consider the cases where k_t and k_r are equal and compare the scenarios with and without the presence of the virtual output queues (VOQ). Each transmitter is only responsible for N/k_t output ports. Likewise, each receiver only takes data from N/k_r input ports. We define $m = N/k_r$, where m is the number of wavelengths in a contention group [11]. Contention only happens when multiple inputs use wavelengths within the same contention group to reach the same receiver. From a practical point of view, instead of port dependent wavelength partition, it is desirable to have the same wavelength partition for the DEMUX rings on each output bus waveguide. Recall the cyclic wavelength routing characteristics of AWGR, we must group different input ports together to form contention groups for different output ports. This static contention group partitioning method may degrade the throughput, but it greatly reduces system complexity [11]. Table 1 illustrated the mapping between the receivers and input nodes for the case of $N = 8$, $k_t = 4$ and $k_r = 4$.

Table 1. The mapping between the input nodes and receivers.

	Output1	Output2	Output3	Output4	Output5	Output6	Output7	Output8
Input1	Rx1	Rx1	Rx2	Rx2	Rx3	Rx3	Rx4	Rx4
Input2		Rx2		Rx3		Rx4		Rx1
Input3	Rx2		Rx3		Rx4		Rx1	
Input4		Rx3		Rx4		Rx1		Rx2
Input5	Rx3		Rx4		Rx1		Rx2	
Input6		Rx4		Rx1		Rx2		Rx3
Input7	Rx4		Rx1		Rx2		Rx3	
Input8		Rx1		Rx2		Rx3		Rx4

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

We assume uniform random traffic with a packet size of 1024B in all the simulations. The inter-arrival time between two packets follows Bernoulli distribution. The line rate is set at 10 Gb/s. We define the maximum offered load as $N \times$ line rate, which does not change with k_t and k_r. For each node, there is a 16-KB input buffer for each transmitter. There is a 10-ns guard time between any two consecutive transmissions due to the ring reconfiguration time [20]. Unlike interconnect switches for data center applications, here we neglect the propagation time from the nodes to the AWGR because it will be less than 10 ps (less than 2 mm distance) for future on-chip multi-core high performance computing systems. The control plane runs a 2-GHz clock and takes three clock cycles to make arbitration for each request. The control plane handles all the contentions based on round robin arbitration. The transmitter will send out the next packet right after it gets the permission while the contended packet will stay in the input buffer and wait for the next grant. The transmitter will send a new request to the control plane immediately after a failed attempt or a successful packet transmission. The case where $k_t = 1$ and $k_r = 1$ represents the simulation for the electrical switch based on input queuing (IQ) crossbar topology.

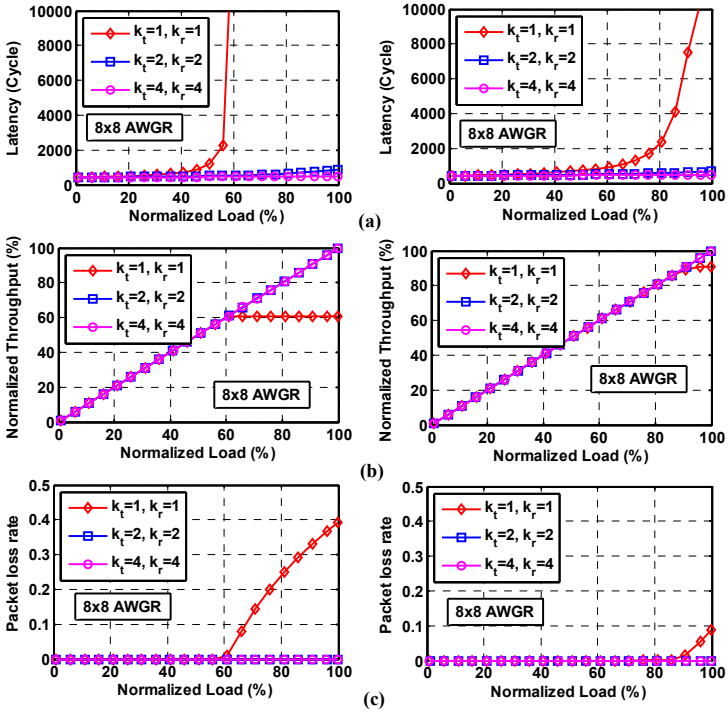


Fig. 3. Performance study on (a) end-to-end latency, (b) throughput and (c) packet loss rate as functions of the offered load for uniform random traffic distribution on proposed architecture with 8 nodes. Left: no VOQ, right with VOQ.

Figure 3 shows the performance study of the proposed architecture based on an 8 × 8 AWGR under various configurations. We investigate the end-to-end latency, throughput and packet loss rate as functions of the input load. Without VOQ, multiple-transmitter/receiver pairs provide significant boost in performance due to the increased statistical multiplexing and the enhanced instantaneous rate at each AWGR inputs and outputs. There is marginal improvement in end-to-end latency when k_t and k_r go from 2 to 4 which indicate even a moderate increase in k_t and k_r from their original value of 1 will substantially improve the performance. The presence of the VOQ greatly improves the system performance for single-transmitter/receiver-pair configuration in all three aspects, but system equipped with multiple transmitter/receiver pairs still performs better especially in terms of end-to-end latency at high

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

(>90%) input load. While the above analysis still holds for the switch with 64 interconnecting nodes, increasing the network size does increase the end-to-end latency, but the influence is insignificant compared to the change of k_t and k_r as shown in Fig. 4. We observe zero packet loss and 100% throughput for all the cases where $k_t \geq 2$and $k_r \geq 2$.

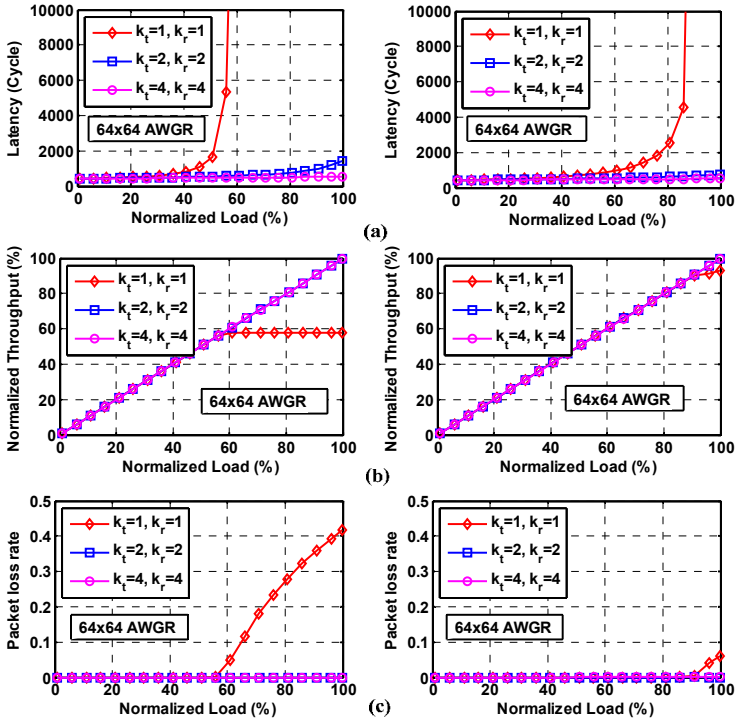


Fig. 4. Performance study on (a) end-to-end latency, (b) throughput and (c) packet loss rate as functions of the offered load for uniform random traffic distribution on proposed architecture with 64 nodes. Left: no VOQ, right with VOQ.

4. Design and characterization of an 8 × 8 switch prototype with $k_t = 4$ and $k_r = 4$

Table 2. OpSIS-IME Process overview.

Parameter/Device Type	Features
Overview	220 nm thick starting Si, 2 µm BOX
	Standard and/or high resistivity handle
	Photonics-only process, no electronics included
Front-end	Two partial etches, one full etch of top silicon
	Six optical implants for modulators and detectors
	100% Ge deposition and implanting
Back-end	Two metal levels, no planarization
	Deep Si Trench for edge coupling
Optical library devices	Grating couplers
	Low loss waveguides (ridge and rib)
	High-speed modulators (reverse-biased pn junction)
	High-speed Ge waveguide photodiodes

The main device consists of three major components, the AWGR, the ring modulators and Ge photo-detectors as shown in Fig. 5(a). The total device area is approx 1.2 mm by 2.4 mm. Figure 5(d) shows a photo of the fabricated device. We do not include the cyclic frequency feature in the 8x8 AWGR design for this particular proof-of-concept demonstration to save device area [16]. Other components such as ring modulators, Ge PDs and edge couplers are from the device library provided by OpSIS-IME [15]. Detailed design parameters for those

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

devices are not included here. Unlike the configuration in Fig. 2, there are no waveguides on top of the ring modulators here as shown in Fig. 5(b). Instead, we couple external light sources into the input bus waveguides through the edge couplers directly, for simplicity. On the output side, as shown in Fig. 5(c), the DEMUX rings select the modulated signals from the output bus waveguides and couples them into the corresponding Ge PDs. The nano-taper edge couplers on the input/output bus waveguides ensure efficient light coupling between the chip and single mode lens fibers. After the mask layout preparation, we use foundry service offered by OpSIS-IME, which runs in a standard microelectronics facility. Table 2 shows the overview of the process offered by OpSIS-IME.

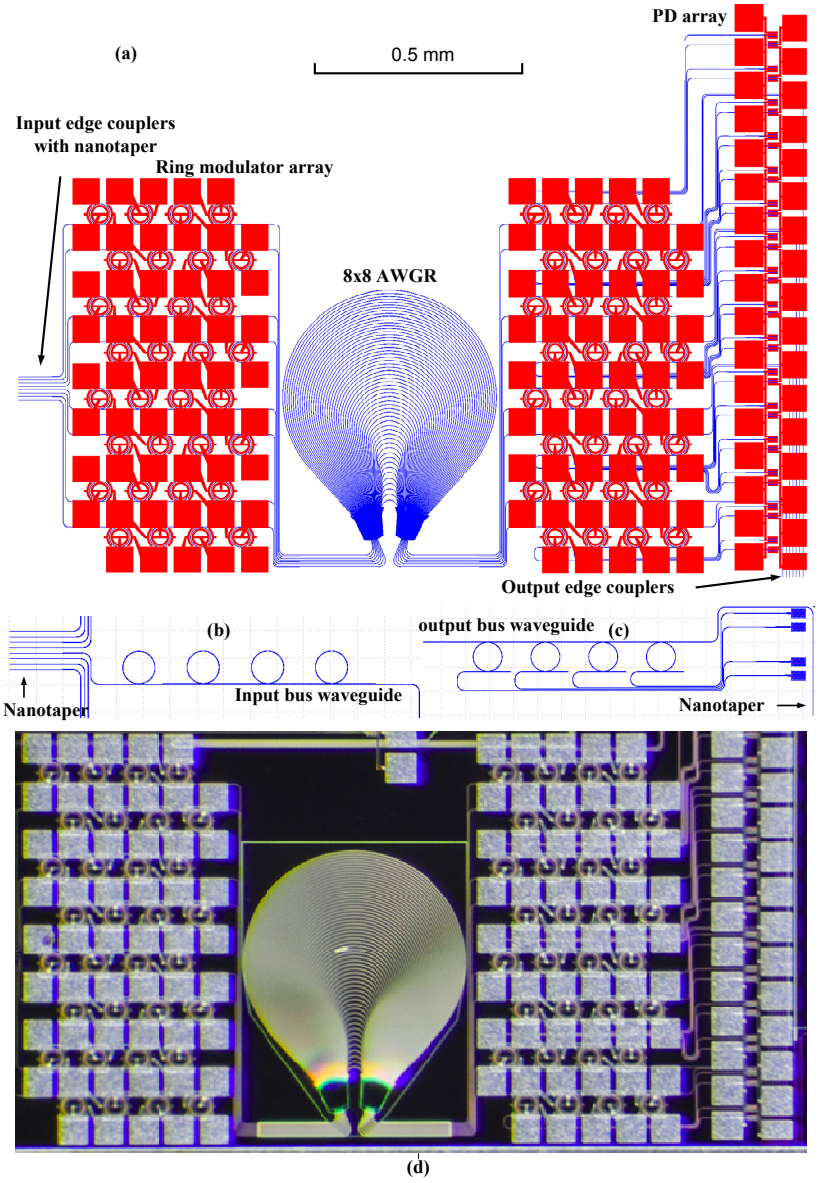


Fig. 5. (a) Device mask layout showing only the waveguide layer and the metal layer (b) rings on one of the input bus waveguides, (c) rings and PDs on one of the output bus waveguides, (d) Photo of fabricated device.

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

The proposed switch can potentially be very compact when compared to the state-of-the-art electronic switch such as Mellanox SwitchX-2 switching IC, which supports up to 64 10-GbE ports and use a 45 mm × 45 mm (2025 mm^2) FCBGA package. We have demonstrated a low-loss 40-port 100-GHz AWG with 3-dB channel bandwidth larger than 10 GHz on the same SOI platform as in [5]. Using an example of the proposed switch based on the 40 × 40 AWGR for comparison based on the recently realized 40-port AWG [5], we estimate as follows. The 40 × 40 AWGR takes approximately 2.0 mm × 2.5 mm, or 5.0 mm^2 [5]. The footprints of the ring and PD are 0.04 mm^2 and 0.03 mm^2, respectively. Assuming we have 2 TX/RX ring pairs on each connecting node (k_t = 2, $\underline{k_r}$ = 2) and one PD for each RX ring, we have 160 rings and 80 PDs in total which occupies 0.04 mm^2 × 160 + 0.03 mm^2 × 80 = 8.8 mm^2 device area. The total area needed for the optical/optoelectronic devices is approx 5.0 mm^2 + 8.8 mm^2 = 13.8 mm^2. Other than the high speed buffers for the optical ring modulators and GE PDs, which are shared by both switches, the proposed switch requires no other high speed electronics leaving sufficient space for the remaining electronics which work at a much lower speed than the line rate. Moreover, the vertical integration of electronic components with optical components can further reduce the overall footprint as illustrated in [21].

Figure 6 shows the measured transmission spectra of the 8 × 8 200-GHz-spaced AWGR, including fiber to chip coupling loss (approx 4 dB per facet) and waveguide propagation loss (approx 2.5 dB/cm). The estimated insertion loss introduced by the AWGR itself is approx 9 dB. The channel cross talk is approx −13 dB, which can be further improved by using special waveguide design in the arrayed arms [22, 23] or including phase error compensating elements in the AWG design [24, 25]. Figure 7(b) and 7(c) shows the mask layout of the depletion-mode ring modulators and Ge photo-detector from OpSIS-IME's library. The measurement results from the testing structures show those devices can support data modulation at 10 Gb/s, which are consistent with the data provided by OpSIS-IME. With a 30-μm radius, the ring FSR is approx 400 GHz. The resonance tunability of the ring is limited to approx 6 pm/V. For this demonstration, we rely on the carrier injection by an optical pump [26] at 1064 nm from above to align the ring resonances with the AWGR passbands. Due to the relative small ring FSR, ring resonances can overlap with multiple AWGR passbands and induce crosstalk. OpSIS-IME now provides ring resonators with FSR up to 1600 THz (8-um ring radius) and resonance tunability larger than one FSR (by electrical carrier injection in a p-i-n junction [15]). This new design eliminates the crosstalk issue in the current design and greatly simplifies the resonance tuning process. Note that, with only one waveguide coupled with each ring as in Fig. 5(b) and Fig. 7(a), the input ring resonators are in all-pass-filter (APF) configurations. This deviation from the original OpSIS-IME design as in Fig. 7(b) degrades the ring on-off extinction ratio under reverse bias. We use carrier injection method instead for data modulation on the input rings. Figure 6(b) illustrates the static transmission spectra of a ring in APF configuration under various forward bias voltages. We observe a 8-dB extinction ratio for data rate up to 0.3 Gb/s which is typical for ring modulator in forward carrier injection mode without pre-emphasis on the electrical driving signal [19]. We will follow OpSIS-IME's original add-drop configuration for future designs in order to guarantee high-speed performance.

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

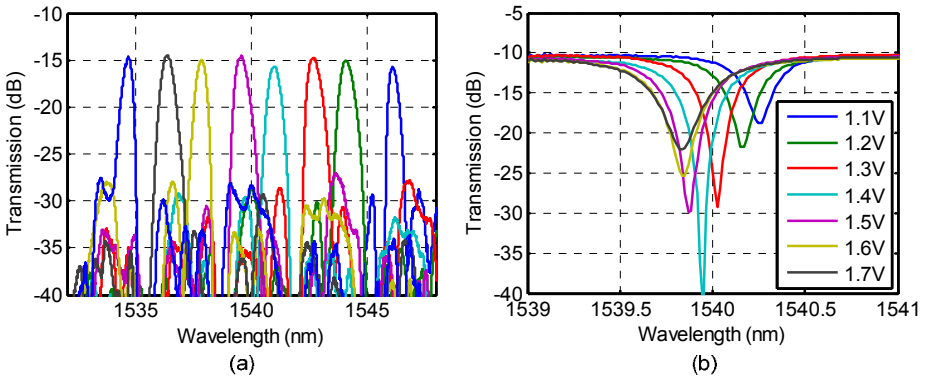


Fig. 6. Measured transmission spectra of (a) the 8x8 AWGR (b) the ring resonator under different forward bias.

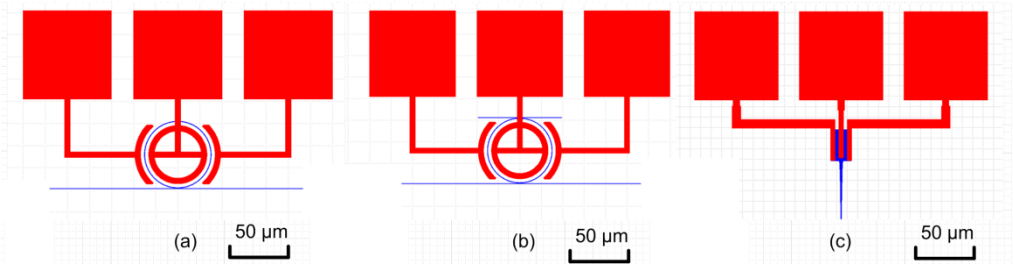


Fig. 7. Layout of the ring modulator (a) in all-pass-filter configuration (without top waveguide), (b) in add-drop configuration (with top waveguide) and (c) Germanium photo-detector (only waveguide and metal layers are shown).

5. Routing experimental setup and results

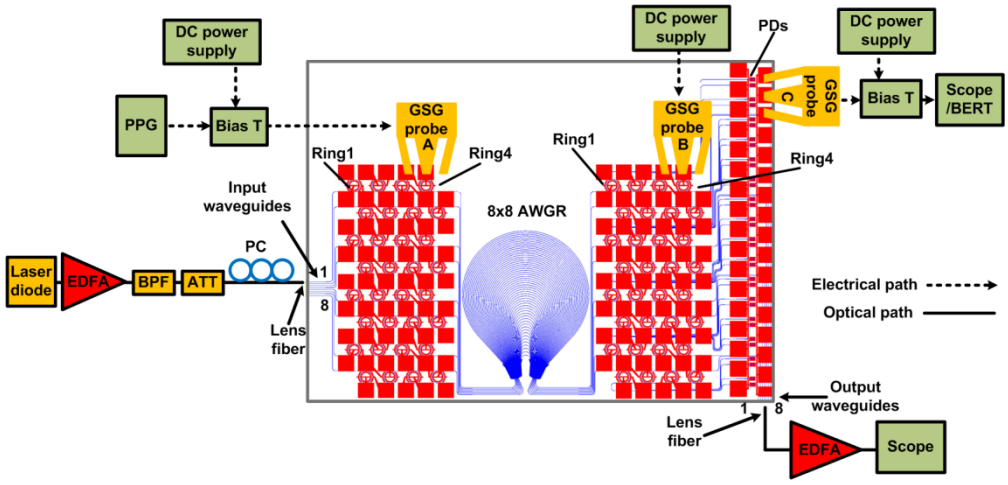


Fig. 8. Experimental setup for the routing demonstration on the fabricated chip. BPF: optical band-pass fiber; ATT: optical attenuator; PPG: pulsed pattern generator.

Figure 8 illustrates the experimental setup for the proof-of-principle routing demonstration on the fabricated chip. The output of a tunable light source enters the switch from one of the input waveguides after amplitude and polarization adjustment. The coupling loss from the lens fiber to the chip is approx 4.0 dB when measured using the straight waveguide test

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

structures on the same chip. We inject the 1064-nm light from two cleaved fibers into the input/output ring pairs from above to align their resonances with the input wavelength and one of the AWGR passbands. A Bias Tee combines the high-speed signal from a Pulse Pattern Generator (PPG) with the DC bias voltage for driving the input ring modulator, while we apply only a DC bias voltage on the output ring. For future design with OpSIS-IME's electrically tunable rings, high speed (>100-Msps update rate) Digital-to-Analog Converters (DACs) will provide the current required to rapidly configure the ring resonance position. With proper alignment between the ring resonance and AWG passband, the modulated light eventually enters the corresponding photo-detector. A second Bias Tee provides a 1-V reverse bias for the PD and extracts the O/E converted high-speed signal for eye diagram and Bit Error Rate (BER) measurements. With the output ring misaligned, a lens fiber collects the optical signal from one of output waveguides and sends the signal to a scope or Optical Spectrum Analyzer (OSA) for monitoring purposes.

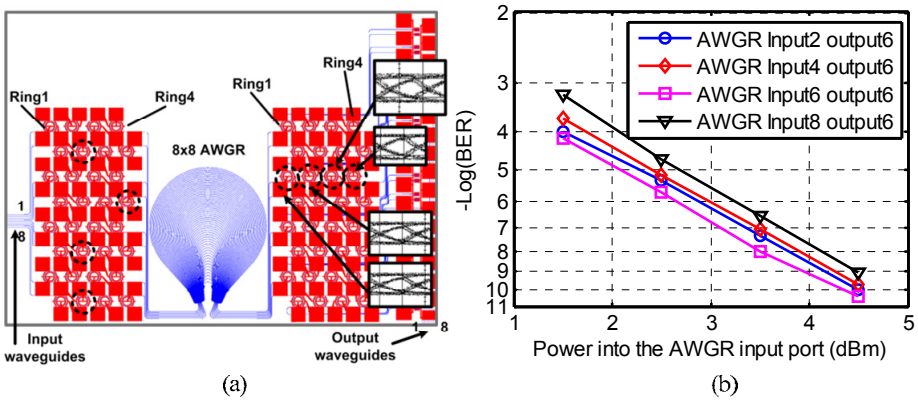


Fig. 9. (a) Experimental configuration for 4-by-1 routing demonstration (b) measured BER.

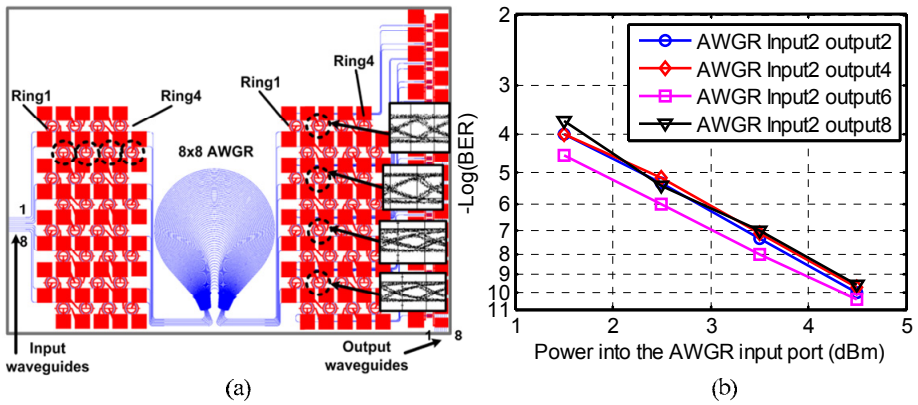


Fig. 10. (a) Experimental configuration for 1-by-4 routing demonstration (b) measured BER.

Figure 9 and Fig. 10 show the selected input/output ring pairs, eye diagrams and corresponding BER measurement for the 1-by-4 and 4-by-1 routing demonstration at 0.3 Gb/s, respectively. The PPG drives the selected input ring through a Bias Tee using a 2^7-1 PRBS sequence. Due to limited resources available at the time of measurement, we only activate one input/output ring pair for each BER measurement. As a result, penalty due to intermodulation crosstalk between multiple active ring modulators is not available at this time. Resonance adjustment on the other rings sharing the same bus waveguide is not

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

necessary since their resonance positions are far away from any of the AWGR passbands. BER measurement shows less than 1-dB power penalty between the different input/output ring pairs, but more than 4-dBm optical power is required to achieve BER below 1E-10 at 0.3 Gb/s. Note that the optical power level is measured before the input lens fiber, so it includes the fiber-to-chip coupling loss (~4dB), AWGR insertion loss (~9dB), waveguide propagation loss and ring coupling loss.

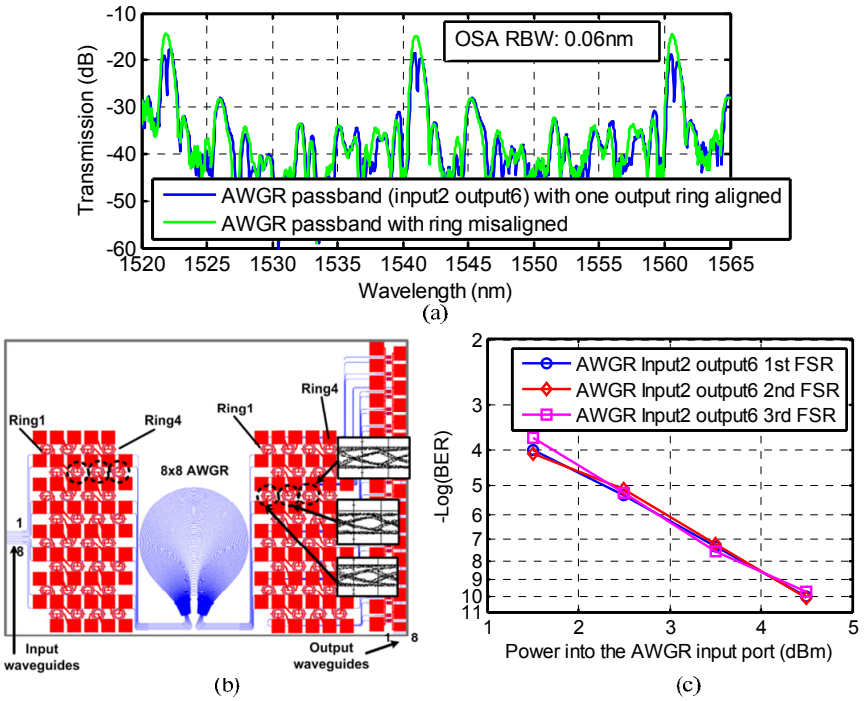


Fig. 11. (a) AWGR transmission spectra over three FSRs, (b) experimental configuration for transmission using multiple FSRs, (c) measured BER.

Figure 11(a) illustrates the measured AWGR passbands across the three FSRs between 1520 nm and 1570 nm on the fabricated device using a broadband ASE from an EDFA and the OSA. The average channel crosstalk is approx −13dB before phase error correction. Using the input/output ring pairs as show in Figure. 11(b), we measure the BER performance of data transmissions on three wavelengths (1521.9 nm, 1540.9 nm and 1560.6 nm) at 0.3 Gb/s. We observe no power penalties as shown in Figure. 11(c). Future high performance computing systems may benefit from parallel-bit transmission by utilizing multiple AWGR FSRs.

6. Conclusion

We propose a scalable chip-scale optical interconnect switch for HPC systems by leveraging unique wavelength routing characteristics of the AWGR, and compact and cost effective device offered by CMOS-compatible silicon photonic integration. We present a comprehensive performance evaluation for the proposed switch, showing low end-to-end latency and high-throughput switching without packet loss even at very high (>95%) input load .We show the performance is impressive even with small number of input/output ring pairs (k_t = 2 and k_r = 2) on each node. Furthermore, we prove the feasibility of the proposed architecture by developing a prototype using silicon photonic integration technology on a SOI platform. We demonstrate successfully wavelength routing functions on the fabricated device.

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013

The development of a more advanced interconnect switch chip with better ring controllability and larger throughput is now in progress.

Acknowledgments

This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.We thank OpSIS-IME for their helpful advice on the device mask layout.

#199251 - $15.00 USD Received 14 Oct 2013; revised 11 Dec 2013; accepted 14 Dec 2013; published 24 Dec 2013